Metabolomics (2008) 4:293–296
DOI 10.1007/s11306-008-0126-2

ORIGINAL ARTICLE

Discriminant Q^2 (DQ^2) for improved discrimination in PLSDA models

Johan A. Westerhuis · Ewoud J. J. van Velzen · Huub C. J. Hoefsloot · Age K. Smilde

Received: 3 June 2008 / Accepted: 15 August 2008 / Published online: 30 August 2008
© The Author(s) 2008. This article is published with open access at Springerlink.com

Abstract In this paper we introduce discriminant Q^2 (DQ^2) as an improvement for the Q^2 value used in the validation of PLSDA models. DQ^2 does not penalize class predictions beyond the class label value. With rigorous Monte Carlo simulations we show that when DQ^2 is used, a smaller effect can be found statistically significant than when the standard Q^2 is used.

Keywords PLSDA · Discrimination · Q^2

1 Introduction

Q^2 is defined as one minus the ratio of the prediction error sum of squares (PRESS) over the total sum of squares (TSS) of the response vector y (Cruciani et al. 1992). When the PLS method was introduced for classification, the Q^2 parameter survived as a measure for class prediction ability and today is regularly used to validate discrimination models such as PLSDA (Lutz et al. 2006; Wiklund et al. 2008). One of the problems of the Q^2 parameters is that it is unclear which Q^2 value corresponds to a good discrimination model. Therefore, the Q^2 value can be compared to a distribution of Q^2 values obtained from models of the same data with randomly permuted class labels Lindgren et al. (1996), Westerhuis et al. (2008). In such a way, statistical significance (P-values) can be obtained for a given discrimination model.

In PLSDA, the response vector **y** consists of class labels -1 and 1 (or 0 and 1) for the two class problem. When the class of new samples is predicted given a PLSDA model (e.g. in a cross validation scheme) the prediction error is summed over all samples to give the PRESS value (see Eq. 1).

$$\text{PRESS} = \sum_i (y_i - \hat{y}_i)^2 \tag{1}$$

Note that here $\hat{y}_i$ represents a real prediction in which sample i was not used in the model building process. The Q^2 value is then calculated as

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} = 1 - \frac{\text{PRESS}}{\text{TSS}}$$

Here the total sum of squares (TSS) is a constant and the PRESS should quantify how well the samples are classified. When the prediction for a sample is close to the discrimination border of 0.5, the PRESS value increases, because the sample is almost predicted wrongly. This seems a good approach. However when a sample of class 1 receives a class prediction of $+1.5$, it is not at all close to the discrimination border, but the PRESS still increases, while it obviously corresponds to a perfect class prediction. As this seems counter intuitive, we developed the discriminant Q^2 (DQ^2) statistic in which the prediction error is disregarded when the class prediction is beyond the class label.

J. A. Westerhuis (✉) · E. J. J. van Velzen · H. C. J. Hoefsloot · A. K. Smilde
Swammerdam Institute for Life Sciences, Universiteit van Amsterdam, Amsterdam, The Netherlands
e-mail: j.a.westerhuis@uva.nl

E. J. J. van Velzen
Unilever Food & Health Research Institute, Vlaardingen, The Netherlands

A. K. Smilde
TNO Quality of Life, Zeist, The Netherlands

$$\underset{\text{Class1}}{\text{PRESSD}} = \sum_{\hat{y}_i < 1} (y_i - \hat{y}_i)^2$$

$$\underset{\text{Class}-1}{\text{PRESSD}} = \sum_{\hat{y}_i > -1} (y_i - \hat{y}_i)^2$$

Thus when the prediction is above 1 for class 1 samples or when the prediction is below −1 for class −1 samples, the prediction error for that sample is ignored. This is represented in Fig. 1 where the red curve represents the prediction error for class 1 samples and the blue curve the prediction error for class −1 samples. It becomes clear that Q^2 penalizes a class prediction of 2 for a class 1 sample in the same way as a class prediction of 0 (which would mean a misclassification). In the DQ^2 statistic, a prediction of 2 for a class 1 sample is disregarded. DQ2 is then defined as

$$DQ^2 = 1 - \frac{\text{PRESSD}}{\text{TSS}}$$

The idea of the DQ^2 statistic is related to other discrimination methods such as logistic regression or SVM in which samples that are close to the discriminating line are more important for the model than samples that are far away from that line.

2 Experimental

1D ^{1}H NOESY NMR spectra of urine samples of 28 male and female human subjects in the age of 35–75 years and mildly hypertensive (Systolic blood pressure: 130–179 mmHg, Diastolic blood pressure: <100 mmHg) were obtained. An exponential window function was applied to the free induction decay (FID) with a line-broadening factor of 0.5 Hz prior to the Fourier transformation. The Fourier transformed NMR data were manually phase and baseline corrected and calibrated against the reference standard TSP resonance at δ 0.0 ppm. The NMR spectra were subdivided into 550 discrete regions ('buckets') of equal width (0.02 ppm), from which the integral regions were determined using AMIX (Analysis of Mixtures, Bruker GmbH, Germany). The spectral region between δ 4.3–5.2 ppm was excluded from the data set to avoid spectral interference of residual water. The urine profiles were normalized to the integral of creatinine methyl peak between δ 3.05–3.10 ppm.

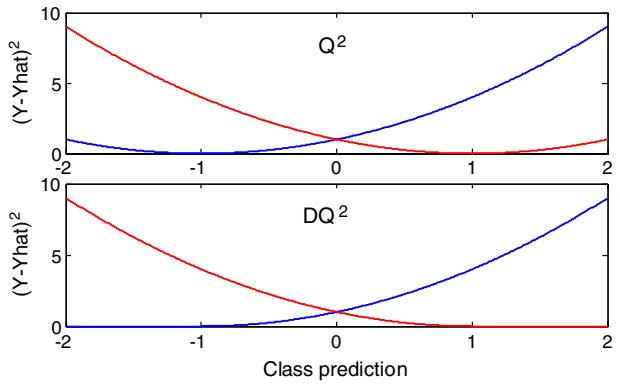

Fig. 1 Representation of the prediction error in Q^2 and DQ^2. The blue curve represents the prediction error for class −1 samples and the red curve represents the prediction error for class 1 samples

Monte Carlo simulations were performed by adding a predefined effect to the spectra of 14 randomly selected volunteers whereas for the other 14 individuals no effect was added. PLSDA was used to discriminate between the two groups. 25 cross model validations [Anderssen et al. (2006)] or sometimes called double cross validation [Smit et al. (2007)] were performed. In each double cross validation the samples were divided into seven groups. For each double cross validation a $(D)Q^2$ value is obtained. Twenty-five double cross validations were performed in which the samples were distributed differently over the seven groups because of the large difference in $(D)Q^2$ value depending on the specific selection of the samples in the seven groups. Thus 25 $(D)Q^2$ values were finally obtained the average $(D)Q^2$ was computed. Then 2,000 permutations were performed in which the class label was randomly permuted and for each permutation again the average $(D)Q^2$ was computed in the same way as described above. The number of average $(D)Q^2$ values of the permutations that are larger than the average $(D)Q^2$ value of the original labeling, divided by 2,000, represents the *P*-value. Finally we repeated the procedure five times with each time a different selection of 14 individuals that received the treatment. In this way five *P*-values were obtained. The average of these five *P*-values is finally used to compare the Q^2 and DQ^2 values.

Two types of effect were added to the 14 selected individuals (see Fig. 2), a univariate effect of a single NMR peak that changed (effect 1) and a multivariate effect (effect 2).

3 Results

Figure 3 shows the results of finding statistical significance for the PLSDA discrimination model between the 14 'treated' individuals and the 14 'non-treated' individuals. In the top figure, the *P*-values decrease from 0.45 to 0 for an increased effect size. The *X*-axis represents how much of the effect was added to the spectrum. Thus the treated spectrum was calculated by the adding the value on the *X*-axis times the effect by the untreated spectrum. Note that in Fig. 2, effect 1 was multiplied by 20 and effect 2 was multiplied by 50 to make them visible. Thus the actually added effects are much smaller than presented in Fig. 2.

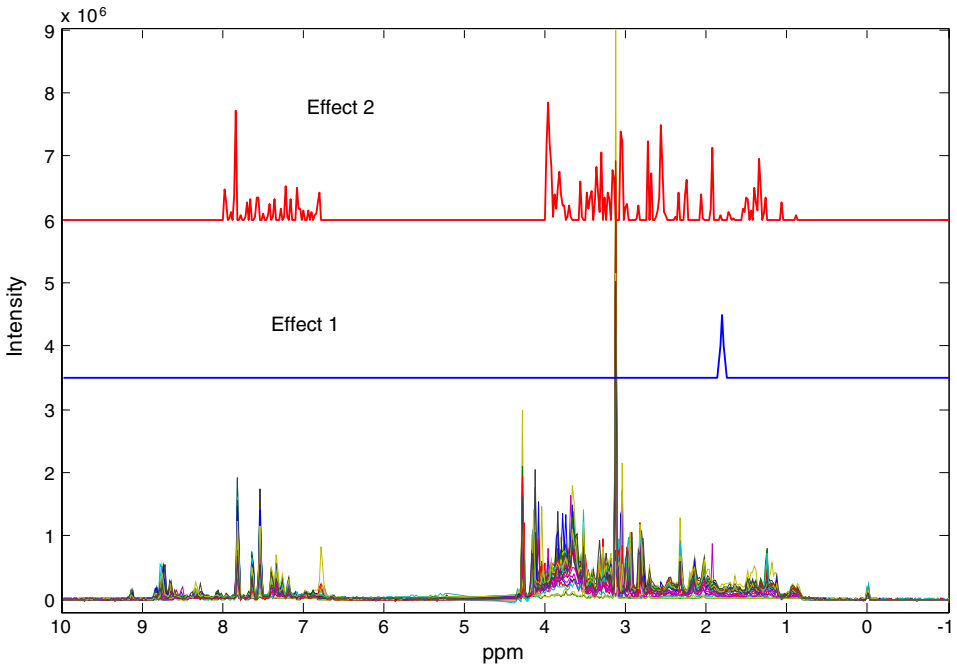

Fig. 2 NMR profiles of urine of 28 healthy individuals with two simulated effects

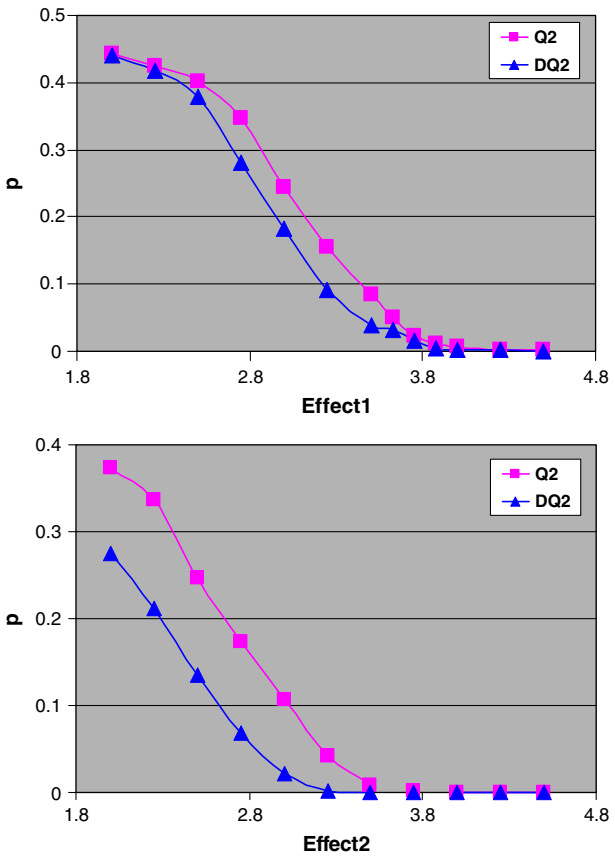

Fig. 3 Statistical significance profiles for PLSDA discrimination model when DQ^2 or Q^2 are used for a univariate effect (top) and a multivariate effect (bottom)

When an $\alpha = 0.05$ significance limit would be used to reject the Null hypothesis of no effect, it can be seen that the effect size needs to be about 3.6 when Q^2 is used while an effect size of about 3.4 already leads to a significant discrimination model when DQ^2 is used. For the multivariate effect (effect 2) in the bottom plot of Fig. 3, the difference between Q^2 and DQ^2 is even larger. Here a multivariate effect size of 2.8 gives a statistical significant model when DQ^2 is used while for Q^2 an effect size of 3.2 is needed to become statistically significant.

4 Conclusion

In this paper the discriminant Q^2 (DQ^2) statistic is introduced as a replacement for the traditionally used Q^2 value to represent class prediction ability. With rigorous Monte Carlo simulation it is shows that statistically significant discrimination models can be found for a smaller effect size when DQ^2 is used than when the traditional Q^2 is used. This is particularly beneficial in metabolomics-based discrimination problems where the biological responses can be subtle and highly variable among the individuals.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Anderssen, E., Dyrstad, K., Westad, F., & Martens, H. (2006). Reducing over-optimism in variable selection by cross-model validation. *Chemometrics and Intelligent Laboratory Systems, 84*(1–2), 69–74.

Cruciani, G., Baroni, M., Clementi, S., et al. (1992). Predictive ability of regression-models.1. Standard-deviation of prediction errors (SDEP). *Journal of Chemometrics, 6*(6), 335–346.

Lindgren, F., Hansen, B., Karcher, W., Sjostrom, M., & Eriksson, L. (1996). Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics, 10*, 521–532.

Lutz, U., Lutz, R. W., & Lutz, W. K. (2006). Metabolic profiling of glucuronides in human urine by LC-MS/MS and partial least-squares discriminant analysis for classification and prediction of gender. *Analytical Chemistry, 78*(13), 4564–4571.

Smit, S., van Breemen, M. J., Hoefsloot, H. C. J., et al. (2007). Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta, 592*(2), 210–217.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81–89.

Wiklund, S., Johansson, E., Sjostrom, L., et al. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry, 80*(1), 115–122.